Information Processing and Management 47 (2011) 856–869

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Learning to select the correct answer in multi-stream question answering

Alberto Téllez-Valero [a], Manuel Montes-y-Gómez [a,*], Luis Villaseñor-Pineda [a], Anselmo Peñas Padilla [b]

[a] Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Sta. María Tonantzintla, Pue. 72840, Mexico
[b] Depto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Juan del Rosal, 16, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 1 April 2009
Received in revised form 26 February 2010
Accepted 21 March 2010
Available online 13 May 2010

Keywords:
Data fusion
Multi-stream QA
Textual entailment
Answer validation

ABSTRACT

Question answering (QA) is the task of automatically answering a question posed in natural language. Currently, there exists several QA approaches, and, according to recent evaluation results, most of them are complementary. That is, different systems are relevant for different kinds of questions. Somehow, this fact indicates that a pertinent combination of various systems should allow to improve the individual results. This paper focuses on this problem, namely, the selection of the correct answer from a given set of responses corresponding to different QA systems. In particular, it proposes a supervised multi-stream approach that decides about the correctness of answers based on a set of features that describe: (i) the compatibility between question and answer types, (ii) the redundancy of answers across streams, as well as (iii) the overlap and non-overlap information between the question–answer pair and the support text. Experimental results are encouraging; evaluated over a set of 190 questions in Spanish and using answers from 17 different QA systems, our multi-stream QA approach could reach an estimated QA performance of 0.74, significantly outperforming the estimated performance from the best individual system (0.53) as well as the result from best traditional multi-stream QA approach (0.60).

1. Introduction

The great amount of available information has motivated the development of different tools for searching and browsing large document collections. The major examples of these tools are information retrieval (IR) systems, which focus on identifying relevant documents for general user queries. Search engines such as Google[1] and Yahoo[2] are a special kind of IR systems which allow to retrieve information from the Web.

It is clear that IR systems have made possible the processing of large volumes of textual information; however, they present serious problems for answering specific questions formulated by users. In fact, they can not properly tackle this task: once the user obtained a list of relevant documents for her question, she still has to review all documents in order to find the desired information. This limitation, along with a growing need for improved information access mechanisms, triggered the emergence of *question answering* (QA) systems. These systems aim at identifying the exact answer to a question from a given document collection. In other words, given an user query in the form of natural language question (e.g., `Which country did Iraq invade in 1990?`), a QA system must detect the text fragment that respond the question (e.g., `Kuwait`) instead of returning a list of documents related to the question words.

* Corresponding author. Tel.: +52 222 2 663 100x8218; fax: +52 222 2 663 152.
E-mail addresses: albertotellezv@ccc.inaoep.mx (A. Téllez-Valero), mmontesg@ccc.inaoep.mx (M. Montes-y-Gómez), villasen@ccc.inaoep.mx (L. Villaseñor-Pineda), anselmo@lsi.uned.es (A.P. Padilla).

[1] http://www.google.com.

[2] http://www.yahoo.com.

0306-4573/$ - see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2010.03.007

Recent research in QA has been mainly fostered by the TREC,[3] CLEF,[4] and NTCIR[5] conferences. These conferences consider different languages (English, European languages and Asian languages respectively) and contemplate different kinds of questions, such as factual questions (e.g., `Where is the Taj Mahal?`), definition questions (e.g., `What is the Quinoa?`), and list questions (e.g., `Who were the members of The Beatles?`). The results from these conferences have shown some interesting facts. On the one hand, they have shown that there are several QA systems that, in spite of their general poor performance, are highly accurate to respond certain kinds of questions. For instance, in the Portuguese QA track at CLEF 2008 (Forner et al., 2008), the system tagged as "diue081" correctly responded 89% of the definition questions, whereas, it only could respond to 35% of the factual questions. On the other hand, these results have also indicated that there is a high complementarity among different QA systems. Just as an example, the combination of the correct answers from all participating systems of the Portuguese QA track at CLEF 2008 (nine systems) could outperform by 49% (reaching a 82% of accuracy) the best individual result for factual questions (55%).

Based on these facts, some advanced approaches known as *multi-stream QA systems* attempt to improve the individual results by taking advantage of the complementarity from existing QA systems. Therefore, the major challenge of this kind of systems is to *select* the correct answer for a given question by combining the evidence from different input systems (or streams). Moreover, considering that for some questions no system is able to extract the correct answer, this challenge also involves determining the cases that require a *nil* response.

Traditional approaches for multi-stream QA rely on measuring the confidence of streams or the redundancy of answers across them (Burger et al., 2002; Clarke et al., 2002; de Chalendar et al., 2002; Jijkoun & de Rijke, 2004; Rotaru & Litman, 2005; Roussinov, Chau, Filatova, & Robles-Flores, 2005). On the other hand, recent approaches consider the application of *textual entailment recognition* (RTE)[6] techniques, which decide about the correctness of answers based on information from a given support text[7] (Glöckner, Sven, & Johannes, 2007; Harabagiu & Hickl, 2006). Motivated by all these previous efforts, in this paper we propose a new kind of *hybrid approach*, which is based on a supervised learning method that combines features from the traditional and textual-entailment approaches. In particular, it considers a set of features that describe the redundancy of answers across streams, the compatibility between question and answer types, as well as the overlap and non-overlap information between the question–answer pair and the support text.

Our experimental results in a set of 190 questions in Spanish language, considering answers from 17 different QA systems,[8] demonstrate the appropriateness of the proposed method. It reached an estimated QA performance of 0.74, significantly outperforming the estimated performance from the best individual system (0.53) as well as the result from best traditional multi-stream QA approach (0.60).

The rest of the paper is organized as follows. Section 2 describes the previous work in multi-stream QA. Section 3 presents our supervised multi-stream QA method, given special attention to the description of the used features. Section 4 shows the evaluation results in a collection of 190 Spanish questions, and compares these results against those from other traditional multi-stream approaches. Finally, Section 5 exposes our conclusions and outlines some future work directions.

2. Related work

A typical QA system consists of three main processes that are carried out in sequence: question analysis, document/passage retrieval, and answer selection. One problem with this kind of architecture is that it is highly affected by cascade errors (Rodrigo, Peñas, & Verdejo, 2008a). In order to reduce this problem new alternative approaches known as *meta QA systems* and *multi-stream QA systems* has been proposed.

On the one hand, meta QA systems internally combine several components or techniques at each QA process. For example, Pizzato and Molla-Aliod (2005) describe a QA architecture that uses several document retrieval methods, and Chu-carroll, Czuba, Prager, and Ittycheriah (2003) present a QA system that applies two different components at each process.

On the other hand, multi-stream QA approaches go a step forward by achieving a superficial combination of several QA systems. Most of these approaches are adaptations of multi-stream techniques from document retrieval (Belkin, Kantor, Fox, & Shaw, 1995; Lee, 1997); nevertheless, in this case, they are mainly focused on *selecting the correct answer* for a given question rather than on ranking all candidate answers. Following, we introduce the traditional approaches for multi-stream QA. In particular, we organize them in five categories taking into consideration some ideas proposed elsewhere (Diamond, 1998; Vogt & Cottrell, 1999). It is important to notice that the first two categories denote *system-centered* approaches, which generate their decisions using information about the system's confidences. Whereas, in contrast, the third and fourth categories correspond to *answer-centered* approaches, which select the final answer exclusively based on its frequency of occurrence across streams.

[3] Text REtrieval Conference; http://trec.nist.gov/.

[4] Cross Language Evaluation Forum; http://www.clef-campaign.org/.

[5] NTCIR Project; http://research.nii.ac.jp/ntcir/.

[6] RTE is determine whether, given two text fragments, the meaning of one text could be reasonably inferred, or textual entailed, from the meaning of the other text (Dagan, Magnini, & Glickman, 2005).

[7] The text fragment from which the answer was extracted.

[8] All these systems were evaluated at the Spanish QA track at CLEF-2006.